Unsupervised Event Extraction from Biomedical text based on Event and Pattern Information

Hong-woo Chun[1], Young-sook Hwang[1], and Hae-Chang Rim[1]

Natural Language Processing Lab., Dept. of CSE,
Korea University, Anam-dong 5-ga, Seongbuk-gu, 136-701, Seoul, Korea
{hwchun,yshwang,rim}@nlp.korea.ac.kr

Abstract. In this paper, we proposed a new event extraction method from biomedical texts. It can extend patterns by unsupervised way based on event and pattern information. Evaluation of our system on GENIA corpus achieves 90.1% precision and 70.0% recall.

1 Introduction

The current electronic revolution taking place via the internet and other networked resources giving easy online access to large collections of texts and data to researchers offers lots of new challenges in the field of automatic information extraction. In genomics, electronic databases are increasing rapidly, but a vast amount of knowledge still resides in large collections of scientific papers such as Medline [1]. In order to extract meaningful information from these data, the study of interactions between genes and proteins is very important.

Most of current approaches usually use predefined patterns of event verbs. However it is impossible for humans to define all the patterns of the event verbs. To make matters worse, there are also insufficient annotated corpora for learning. Thus, our proposed method is to use not only the patterns of event verbs but also statistical measures. These measures determine the frequency of verbs, nouns and their co-occurrence information, and the dependency relation. As a result, we show the ranking of events and patterns, thereby allowing us to extract reliable events.

In this paper, *Entity* is a biomedical class name such as protein, gene, cell, tissue, etc. *Event* is defined as the binary relation between subject entity and object entity for special event verbs.

2 System Architecture

The ultimate goal of our research is to build a network of gene or protein interaction using events extracted from biomedical texts. An illustration of the relevant portion of the architecture is shown in Figure 1.

In order to extract reliable events, we need a preprocessing procedure[Figure 1]. The first module is POS tagging and chunking. These processes transform